The Statistical Perspective

Statistical terminology

In statistics, you'll also see the data described in terms of **independent variables** and **dependent variables**. These names come from the idea that the value of one variable may *depend* on the value of some other variables. For example, the selling `price` of a house is the dependent variable that *depends* on some independent variables—like the house's `location` and `size`.

In the example of clothing products we looked at earlier in this lesson:

SKU	Make	Color	Quantity	Price
908721	Guess	Blue	789	45.33
456552	Tillys	Red	244	22.91
789921	A&F	Green	387	25.92
872266	Guess	Blue	154	17.56

We might use data in each row (e.g. `(908721, Guess, Blue, 789, 45.33)`) to predict the sale of the corresponding item. Thus, the sale of each item is *dependent* on the data in each row. We can call the data in each row the independent variables and call the sale the dependent variable.

Input and output

From a statistical perspective, the machine learning algorithm is trying to learn a hypothetical function `(f)` such that:

```
Output Variable = f(Input Variables)
```

Typically, the *independent variables* are the input, and the *dependent variables* are the output. Thus, the above formula can also be expressed as:

```
Dependent Variable = f(Independent Variables)
```

In other words, we are feeding the independent variables into the function, and the function is giving us the resulting values of the dependent variables. With the housing example, we might want to have a function that can take the independent variables of `size` and `location` as input and use these to predict the likely selling `price` of the house as output.

Yet another way to represent this concept is to use shorthand notation. Often, the input variables are denoted as `X` and the output variable is denoted as `Y`:

```
Y = f(X)
```

In the case of multiple input variables, X would be an **input vector**, meaning that it would be composed of multiple individual inputs (e.g. `(908721, Guess, Blue, 789, 45.33)`). When this is the case, you'll see the individual inputs denoted with a subscript, as in X_1, X_2, X_3, and so on.

NEXT